■ 論 文 ■

대중교통 OD구축을 위한 대중교통카드 데이터의 오류와 결측 분석 및 보정에 관한 연구

The study on error, missing data and imputation of the smart card data for the transit OD construction.

박 준 환
(前서울시정개발연구원 부연구위원)

김 순 관
(서울시정개발연구원 선임연구위원)

조 종 석
(한국교통연구원 책임연구원)

허 민 욱
(서울시 도로교통시설담당관)

목 차

Key Words : 대중교통, 카드 데이터, 데이터 오류, 결측 보정, 대중교통 OD
transit, card data, data error, imputation, transit OD

요 약

대중교통 교통카드 도입 이후, 점차 이용율이 증가되고 있다. 카드 데이터를 통해 얻을 수 있는 자료를 고려할 때 대중교통 카드 이용의 증가는 통행패턴 분석 및 정책적 측면에서 중요한 의미를 가지고 있다. 그 중에서 특히 존별 대중교통 통행수요(O/D)를 손쉽게 파악할 수 있다는 점에서 높은 중요성을 가진다.

카드데이터를 통해 대중교통 존별 통행수요(O/D)를 파악함에 있어서 데이터 자체의 오류에 대한 분석이나 결측에 대한 보완 과정이 반드시 필요하다. 본 연구에서는 반드시 선행되어야 할 과제이지만 아직 연구사례가 없었던 카드데이터의 오류와 결측에 관해 살펴보았다.

그 결과, 통행수요(O/D)분석과 관련한 오류나 결측에 대한 특성을 제시하였고, 결측에 대한 보정방안을 제안하였다. 그리고 제시된 결측방안들에 대한 적용 및 평가와 함께 활용방안을 제시하여, 향후 보다 신뢰성있는 대중교통 OD구축을 위한 기반을 마련하였다.

The number of card users has grown steadily after the adaption of smart card. Considering the diverse information from smart card data, the increase of card usage rate leads to various useful implications meaning in travel pattern analysis and transportation policy. One of the most important implications is the possibility that the data enables us to generate transit O/D tables easily.

In the case of generating transit O/D tables from smart card data, it is necessary to filter data error and/or data missing. Also, the correction of data missing is an important procedure. In this study, it is examined to compute the level of data error and data missing, and to correct data missing for transit O/D generation.

Ⅰ. 서론

1. 연구의 배경 및 목적

2004년부터 통합 대중교통 거리비례 요금제가 시행됨에 따라 대중교통 교통카드의 이용이 증가되었다. 이러한 대중교통 카드 이용은 이용자의 편의를 제공해주는 측면에서도 중요하지만 카드 데이터를 통해 얻을 수 있는 자료를 고려할 때 교통분석 및 정책적 측면에서 중요한 의미를 가지게 되었다.

대중교통 카드를 통해 얻을 수 있는 정보 중에 가장 중요한 것은 대중교통 통행수요를 전수화 조사에 가깝게 파악할 수 있다는 것이다. 기존에 가구통행실태조사 등과 같이 대규모 조사를 통해서만 파악할 수 있었던 존별 대중교통 통행수요(O/D)를 대중교통 카드데이터를 통해 손쉽게 파악할 수 있게 된 것이다.

그러나 카드데이터를 통해 대중교통 존별 통행수요(O/D)를 파악하기 위해서 선행되어야 할 연구가 카드데이터에 대한 신뢰성을 파악하는 것이다. 대부분의 대규모 데이터들이 그러하듯이 카드데이터에서도 여러 가지 오류가 포함되어 있고, 결측된 데이터도 포함되어 있다. 이러한 오류와 결측에 대해 정확한 인식과 보정방안이 마련되어야 보다 정확한 대중교통 통행수요에 대한 파악이 가능할 것이다. 현재로서는 대중교통 카드데이터가 가지고 있는 오류나 결측에 대한 인식이나 발생 정도에 대한 연구도 없는 상태라는 점에서 본 연구가 더욱 의의를 가진다고 판단된다.

본 연구에서는 이러한 연구의 배경과 필요성을 바탕으로 교통카드데이터를 이용하여 대중교통 통행수요(O/D)를 구축함에 있어서 요구되는 오류 및 결측에 대해 분석한다. 더불어 이러한 오류 혹은 결측을 보정하는 방안 제시를 본 연구의 목적으로 한다.

2. 연구의 절차 및 방법

본 연구는 크게 4단계를 통해 수행한다. 첫째, 카드데이터의 정의 및 이용 현황, 포함되어 있는 정보의 종류 등에 대해 요약한다.

둘째, 카드데이터가 가지고 있는 오류 및 결측을 인식하기 위한 과정을 수행한다. 즉, 오류의 종류와 특성에 대한 분석 수행과 함께 결측 데이터의 발생빈도 패턴 등에 대해 살펴본다. 결측데이터의 발생 패턴을 수단별, 시간별, 요일별로 분석하여 제시한다.

셋째, 결측자료를 보정하는 방안을 개발하여 제시한다. 이 방법은 크게 개인통행에 대한 정보를 이용하는 방법과 노선별 승하차 패턴을 이용하는 방법으로 구성된다. 그리고, 구축된 방법을 카드데이터에 적용 및 결과도출을 통해 결측 보정 방법론을 검증한다.

본 연구는 이러한 과정을 거쳐 카드데이터를 이용한 통행수요 분석에서 요구되는 신뢰성있는 데이터의 구축 방안을 제시한다.

Ⅱ. 카드데이터 이용현황 및 특성 분석

1. 교통 카드 정의 및 이용현황

교통카드는 교통수단에의 탑승 및 이용을 위해 사용되는 전자화폐의 하나로서 현금, 수표, 신용카드 등 기존의 화폐와 동일한 가치를 지니는 카드로 정의할 수 있다.

서울시의 교통카드 이용률은 2004년 7월 대중교통체계 개편 이후 지속적으로 증가추세를 보이고 있다. 2005년 버스 교통카드 이용률은 2004년 대비 7.3%가 증가한 90.2%이며, 지하철은 4.7% 증가한 72.2%, 대중교통 전체로는 6.6% 증가한 82.1%이다. 지하철의 경우 회수권(MS)을 이용하는 비율로 인해 버스에 비하여 낮은 카드이용률을 나타내고 있다. 이렇게 카드 이용률이 증가함에 따라 카드데이터의 활용성은 높아지고, 데이터의 신뢰성이 더욱 중요해지고 있다.

〈표 1〉 연도별 대중교통 카드이용률 (단위:%)

구분	2003년	2004년	2005년
버스	77.3	82.9	90.2
지하철	63.5	67.5	72.2
대중교통 전체	70.8	75.3	82.1

2. 대중교통 카드데이터의 정의 및 특성

대중교통 카드는 다양한 정보를 저장하고 있다. 출발 정류장 및 도착정류장을 비롯하여 19가지의 정보로 구성된 카드데이터 정보를 정리하면 〈표 2〉와 같다.

이러한 카드데이터를 통해 다양한 교통정보를 얻을 수 있지만 가장 중요한 점은 존별 통행수요(O/D)에 대

〈표 2〉 교통카드 데이터의 정보 구성

컬럼명	예제	비고
카드ID	36eOB01Z	랜덤ID
승차일시	20061101000204	2006. 11. 1. 0시 2분 4초
교통수단ID	120	지선버스
환승횟수	1	환승횟수
버스노선ID	11110246	-
사업자ID	111002600	군포교통㈜
차량ID	111742849	-
사용자 구분	1	일반,어린이,청소년
운행출발일시	20061031230342	2006년 10월 31일 23시 3분 42초
승차정류장ID	9695	정류장코드
하차일시	20061101001101	2006년 11월 1일 0시 11분 1초
하차정류장ID	9484	정류장코드
이용객수	1	탑승객수
승차금액	0	승차 시 요금
하차금액	100	하차 시 요금

〈표 3〉 가구통행실태조사와 카드데이터의 O/D 특성 비교

항 목	가구통행 실태조사 OD	카드데이터 OD
조사지역	수도권 및 조사대상지역	교통카드 단말기 설치 도시
조사시간	특정 조사일	상시조사 가능
조사비용	고가	저가
오차 원인	전수화 및 입력오차	시스템의 오류 및 결측 오차
통행목적	조사가능	조사불가
통행수단	모든 수단	대중교통
자료 취득시간	장시간 소요됨	단시간 소요됨
사고로 인한 조사의 신뢰도	영향을 받음	영향 없음
통행대상	수도권 및 조사대상지역 내 모든 통행인구	현금 승차자 외 교통카드이용자

한 정보를 손쉽게 구축할 수 있다는 것이다. 지금까지 O/D는 가구통행실태조사를 통해서 구축되어 왔는데, 이러한 가구통행실태조사에서의 O/D와 교통카드를 통해 얻을 수 있는 O/D는 〈표 3〉에서 제시하는 바와 같이 서로다른 특성을 가지고 있다.

카드데이터는 가구통행실태조사에 비해 여러 가지 장점을 가지고 있다. 즉, 카드데이터 시스템이 구축된 이후에는 저렴한 비용으로 상시적인 O/D 구축이 가능하고, 대중교통 수단별 O/D와 함께 환승을 포함한 Trip chain 분석이 용이하다는 점이 대표적 장점이 될 것이다. 더불어, 저렴하고 비교적 실시간 O/D의 구축이 가능하므로 GIS등과의 연계를 통한 공간적 분석과 같은 2차 데이터 분석·연구가 용이하다는 점도 카드데이터 O/D 구축의 장점이라고 할 수 있다.

그러나 카드데이터는 데이터의 오류 및 결측에서 오차가 발생할 수 있다는 점에 주목할 필요가 있다.

따라서 본 연구에서는 카드데이터를 통해 O/D를 구축함에 있어서 반드시 고려되어야 하는 쟁점인 데이터의 오류 및 결측에 대해 살펴보고 보정방안을 제시하고자 한다.

Ⅲ. 카드 데이터의 오류 및 결측 분석

1. 오류의 정의 및 특성 분석

교통카드 데이터를 이용하여 O/D를 구축할 경우, 여러 가지 장점들이 존재하지만 반드시 거쳐야하는 과정이 있다. 수집된 교통카드 데이터에 포함되어 있는 오류에 대해 정의·분석한 후, 보정하는 것이다. 본 연구에서는 오류 분석을 위해 2007년 3월 29일(목) ~ 4월 4일(수)까지의 데이터를 통해 검토하였다.

이 때 데이터 오류라 함은 데이터가 수집되었으나 수집된 데이터의 일부 혹은 전체에서 자체적인 모순 또는 논리적인 오류를 내포하고 있는 경우를 뜻한다.

본 연구에서는 카드데이터의 오류를 크게 두종류로 구분하여 살펴보았다. 첫째는 데이터의 특정 항목에 해당하는 값이 합리적인 기대치를 벗어나는 기대치 오류이며, 다른 한 가지는 둘 이상의 항목이 논리의 흐름 상 모순을 일으킴으로 발생하는 논리 오류이라고 할 수 있다.

기대치 오류를 통해 검토해본 오류 내용은 다음과 같다.

① 탑승객수가 "0"으로 입력된 경우

② 첫 승차의 기본요금이 250원 이하로 입력된 경우 및 1800원을 초과

이 외에도 등록된 교통수단 이외의 교통수단이 입력된 경우, 환승회수에 "0"부터 "4"까지의 값 이외의 값이 입력된 경우, 사용자구분 코드가 1(일반), 2(초등학생), 4(청소년) 이외의 값 입력된 경우 등 다양한 오류들을 검토하였으나 조사된 자료에서 발견된 오류는 위의 두가지 오류였다.

이에 반해 논리적 오류는 다음과 같은 경우를 발견할 수 있었다.

① 하차시간이 승차시간보다 빠른 경우

〈표 4〉 교통카드 데이터의 오류 비율

구분	오류유형	오류 발생률
기대치 오류	① 탑승객수	0.001%~0.0001%
	② 기본요금	0.02% ~ 0.03%
논리 오류	① 승하차시간 관련	0.004% ~ 0.007%
	② 승하차시간 관련	0.02% ~ 0.03%
	③ 승하차장 동일	2.9% ~ 3.1%

② 하차와 승차시간의 시간차가 3시간 이상인 경우

일주일 자료의 분석 후 매일 발생하는 각 오류 종류별 발생률을 정리하면 〈표 4〉와 같다.

분석 내용 중 승차정류장과 하차정류장이 동일한 경우도 있었다. 이러한 경우는 버스노선을 오인하여 승차 후 바로 하차하는 경우, 승차와 하차정류장이 가까워 승차 후 바로 하차 태그(Tag)를 하는 경우 등을 생각할 수 있다. 이러한 문제는 명백한 오류라고 단정하기도 어려울 뿐 아니라 소존인 행정동 단위의 O/D분석에 있어서 존 내에 위치하는 가까운 정류장의 하차 정류장 오차는 중요한 문제가 아니라고 판단하여 본 연구 범위에서는 제외하였다. 단, 발생빈도에 대해서는 〈표 4〉에서 다른 오류들과 함께 제시하였다.

승하차동일 문제를 제외하면 본 연구에서 발견된 오류의 발생률은 0.1%미만으로 아주 미미하다. 따라서 별도로 보정방안을 마련하는 것은 불필요한 것으로 보이지만, 통행패턴을 분석하는 경우에 이러한 오류 데이터는 이상치로 작용할 수 있으므로 오류데이터의 확인 및 제거 작업이 필요하다. 즉, 이러한 오류에 대한 필터링 작업을 통해 보다 신뢰성있는 교통수요 데이터의 구축 및 통행패턴 분석이 가능하다.

2. 결측데이터의 정의 및 패턴 분석

교통카드 데이터를 이용하여 신뢰도 높은 O/D를 구축하기 위해서는 결측 데이터에 대한 분석 및 보정의 절차가 필요하다.

이 때, 결측 데이터라 함은 탑승객의 교통카드 미접촉 및 통신두절 등의 원인으로 데이터의 일부가 불완전하게 수집된 데이터를 뜻한다. 불완전한 정보를 완전한 자료로 가공하기 위해서는 결측 데이터에 대해 보정 작업을 거쳐야 한다. 특히 카드데이터를 이용하여 OD를 구축함에 있어서 승차나 하차의 결측은 출발지나 도착지를 알 수 없게 되므로 결측을 보정할 수 있는 방법이 반드시 필요하다.

지하철의 경우, 대부분 하차 시에도 교통카드를 단말기에 접촉하여 하차결측이 문제가 되지않으므로 본 연구에서는 버스를 대상으로 한 하차결측에 초점을 맞추어 연구를 수행하였다. 이 때, 승차 및 하차 시 교통카드를 단말기에 접촉하지 않아서 발생하게 되는 데이터의 결측을 승차 및 하차결측이라 정의한다. 승차 결측은 2007년 3월 10일(토)~3월 23일(금)의 데이터를 분석한 결과 전혀 없었기 때문에 본 연구에서는 하차결측을 중심으로 분석을 수행한다.

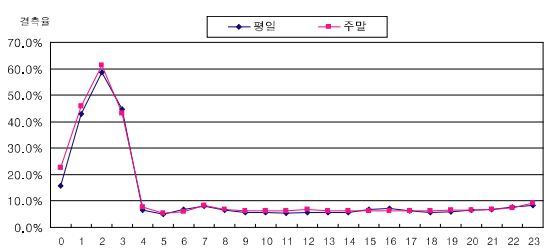
1) 하차결측 발생량 분포

3월 10일(토)~23일(금)의 데이터를 바탕으로 하차정보의 결측분포를 분석하였다. 하차정보 결측 데이터의 총량적 분석 결과는 〈표 5〉와 〈그림 1〉과 같다.

전일 통행을 세분화하여 시간대별 및 교통수단별 결측률의 분포를 분석하였다. 우선, 결측률의 시간대별 분포를 살펴보면 밤 10시부터 새벽 4시까지의 결측률이 매우 높게 나타났다. 이는 심야시간대에 운행하는 버스는 주로 하차 시 교통카드를 찍을 필요가 없는 광역버스이기 때문인 것으로 판단된다.

〈표 5〉 하차정보 결측량 (단위 : 통행, %)

날짜	하차정보 결측량	총 통행량	비율
3월10일(토)	532,375	8,636,885	6.2%
3월11일(일)	400,724	6,058,217	6.6%
3월12일(월)	760,966	11,629,656	6.5%
3월13일(화)	782,278	11,776,441	6.6%
3월14일(수)	796,040	11,987,042	6.6%
3월15일(목)	778,580	11,686,085	6.7%
3월16일(금)	788,268	11,810,197	6.7%
3월17일(토)	681,546	9,815,499	6.9%
3월18일(일)	447,806	6,752,693	6.6%
3월19일(월)	758,734	11,646,728	6.5%
3월20일(화)	769,194	11,686,948	6.6%
3월21일(수)	762,106	11,432,311	6.7%
3월22일(목)	773,433	11,623,852	6.7%
3월23일(토)	796,322	11,886,586	6.7%



〈그림 1〉 시간대별 하차정보 결측 비율

그래프 비교 결과, 평일과 주말의 하차정보 결측 비율은 시간대별로 거의 차이가 없는 것으로 보인다.

2) 버스 수단별 하차결측 발생분포

버스의 수단별 하차정보 결측데이터의 분석 결과는 다음과 같다. 버스를 마을, 간선, 지선, 광역버스로 구분하여 살펴보았다.

광역버스 83.2%, 간선버스 12.4%, 지선버스 11.4% 순서로 하차정보의 결측률이 높은 것으로 분석된다. 이는 서울시계 내외를 오가며 운행하게 되는 광역버스는 환승할인 혜택이 없기 때문에 대부분 하차시 단말기 접촉을 하지 않는 것으로 풀이된다[1]. 새벽 시간대 하차정보 결측률이 높은 것도 새벽 시간대에 주로 운행하는 광역버스의 결측률이 높음에 기인한 것으로 설명할 수 있다.

〈표 6〉 수단별 하차정보 결측 (단위 : 천통행, %)

구분	하차정보 결측수 / 총 통행량			
	마을버스	간선버스	지선버스	광역버스
평일	89	288	291	76,121
평균	1,165	2,304	2,551	14,036
(6.6%)	7.7%	12.5%	11.4%	84.4%
주말	64	194	203	48,969
평균	808	1,619	1,775	9,976
(6.7%)	8.0%	12.0%	11.5%	79.1%
전체	82	259	264	67,695
평균	1,054	2,092	2,310	12,776
(6.6%)	7.8%	12.4%	11.4%	83.2%

3) 노선별 결측발생 분포

하차결측 정보가 시간대나 요일별로는 크게 차이가 없는 것으로 나타났다. 그러나 시간별 차이와는 별도로 결측의 공간적 분포에 대한 검토도 필요하다. 특히 대중교통 통행수요의 공간적 분포를 제시하는 O/D 분석에 있어서 결측자료의 공간적 분포 확인은 중요한 작업이다. 이러한 결측의 공간적 분포는 노선별 결측 분포를 통해 알 수 있다.

분석 결과 결측의 분포가 버스 종류별, 노선별로 차이가 있는 것으로 나타났다. 그 예로 〈표 7〉에서 각 버스 수단별 노선들에 대한 결측정보 비율을 제시하였다. 이 표에서 나타난 바와 같이 약1.5%에서부터 50% 이상의 다양한 결측 분포를 보여주고 있다.

이는 결측분포가 공간적으로 균일하지 않은 것을 의미하며, 공간적 분포에 대한 인식이 중요한 통행수요(O/D)분석에 있어서 노선별로 세분화된 결측보정이 반드시 필요하다는 것을 의미한다.

〈표 7〉 노선별 하차정보 결측 (단위:%)

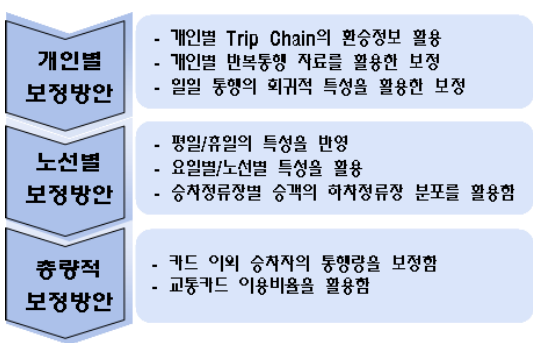
구분	버스번호	결측율	구분	버스번호	결측율
마을 버스	강동01번	24.61	지선 버스	7728번	35.51
	강동05번	22.42		706번	30.84
	마포04번	2.15		1164번	2.43
	성북15번	1.58		1116번	1.94
간선 버스	540번	27.54	순환 버스	탄천01번	51.74
	407번	27.29		02번	13.43
	102번	6.44		62번	7.33
	120번	6.41		61번	6.23

Ⅳ. 결측자료 보정 방안 구축

1. 개요

지금까지 카드데이터의 오류 및 결측정보에 대해 살펴보았다. 그 결과 오류는 발생빈도가 드물어 통행수요나 패턴 분석에 미치는 영향이 미미하고, 보정보다는 필터링을 통한 삭제가 합리적일 것으로 판단된다. 그에 반해 하차정보 결측은 발생정도도 빈번하고, 지역별, 시간대별, 수단별 분포도 균일하지 않기 때문에 단순 삭제로 처리할 수 있는 문제가 아니라고 판단된다.

따라서 본 연구에서는 하차정보 결측을 보정할 수 있는 방안을 제안하고, 그 효과를 검증하여 제시함으로써



〈그림 2〉 하차정보의 결측에 따른 보정방안

1) 2007년 7월 1일 이후부터 수도권 광역버스도 환승할인이 되므로 점차하차결측이 줄어들 것으로 보인다.

향후 카드데이터를 이용한 통행수요 분석에서 유용하게 활용할 수 있도록 한다.

본 연구에서 제안하는 하차결측 보정방안은 크게 3단계로 구성된다.

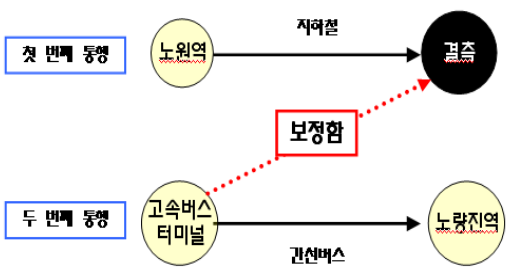
각 단계의 모든 절차를 수행할 수도 있지만, 경우에 따라 특정한 방법만 적용할 수도 있다. 다음 절에서 각 단계의 구체적인 방법론과 적용성을 설명한다.

2. 개인별 통행자료 활용을 통한 보정

개인별 통행정보를 분석하여 통행사슬(trip chain)속에서 반복되는 패턴을 찾아 결측자료를 보정하는 방법이다. 이 방법은 3가지 세부적인 방법으로 구분할 수 있다.

1) 환승 정보 활용 방안

개인의 환승 정보를 이용한 보정 방법이다. 즉, 결측된 하차 정보 후 다음의 통행이 존재하는 경우, 다음 통행의 출발지와 가장 가까운 이전 통행의 하차 정류장을 결측된 정류장으로 보정하는 방법이다.



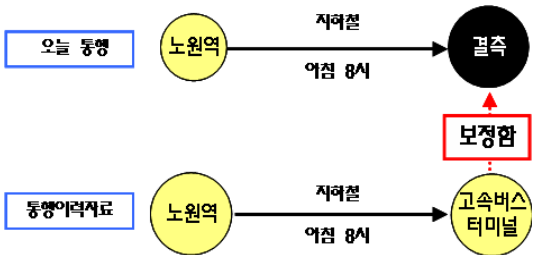
〈그림 3〉 환승 정보를 활용한 보정 방안

예를 들어 한 개인이 노원역에서 승차하여 지하철로 이동하는 통행에 있어서 하차정보없이 환승통행으로 "고속버스터미널"에서 버스를 탄 경우, 첫 번째 통행의 하차지점을 "고속버스터미널역"으로 보정할 수 있다는 것이다.

이 방법은 결측된 하차정보 이후에 연속된 통행이 존재하는 경우에 한하여 적용가능하다.

2) 반복되는 개인별 통행 자료 활용

통행이 여러 날짜에 걸쳐 반복되는 개인별 통행특성을 경우에 이용할 수 있는 보정 방법이다. 예를들어 한 개인이 매일 오전8시경에 노원역에서 고속버스터미널로 출근하는 통행이 반복되는 특성을 가졌다는 사실을 발견하는 경우, 수집된 이력 자료를 이용하여 결측정보를 완전한 정보로 보정해 주게 된다.

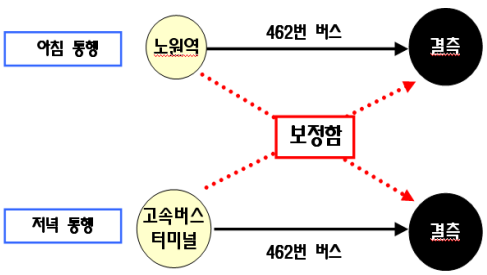


〈그림 4〉 개인의 반복통행을 활용한 보정

3) 일일통행의 회귀적 특성 활용 방안

개인의 하루 통행은 집에서 출발하여, 집으로 돌아온다는 개념에 근거한 보정 방법이다. 하루 중 대응되는 특성을 나타내는 두 개의 통행은 서로 보정이 가능하다.

특정 개인의 당일 통행 중 두 건의 하차 정보 결측이 발생하였는데, 두 통행이 상호 대응되는 특성을 나타내는 경우가 있다. 즉, 같은 노선과 동일한 요금, 승차시간이 아침과 저녁으로 대응되는 경우는 두 가지 통행이 상호 연관된 것으로 판단할 수 있다.



〈그림 5〉 개인 통행의 회귀적 특성을 이용한 보정

〈그림 5〉에서 제시한 예시의 경우, 아침 통행의 결측정보를 "고속버스 터미널"로 보정하고, 저녁 통행의 결측정보를 "노원역"으로 보정할 수 있는 것이다. 이러한 보정 방법은 상호 대응되는 통행 중 어느 한 쪽의 통행만 결측이 발생한 경우도 활용 가능하다.

4) 적용성 검토

개인별 이력 자료를 이용한 보정방안은 각 개인의 통행특성을 반영하여 보정한다는 점에서 신뢰성을 기대할

수 있다는 장점이 있다. 그러나 이 방법론의 적용에 있어서 몇가지 문제점을 가지고 있다.

첫째, 평일 하루평균 약1,100만 통행이 기록되는 카드데이터에서 각 방법론에 적용가능한 개인별 이력데이터를 분석하는 것은 상당한 시간적·하드웨어적 비용이 수반되어야 한다는 단점이 있다.

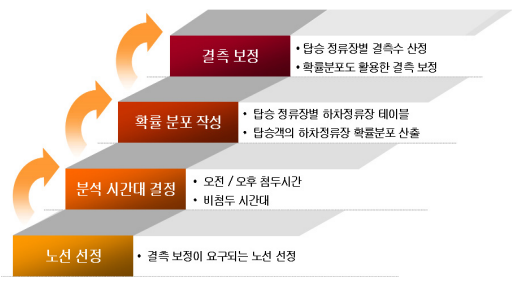
둘째, 개인의 통행에 있어서 반복되는 패턴이 존재해야 한다. 다시말해, 개인의 통행 이력자료에 있어서 일정한 패턴이 존재(예를 들어 일정한 시간에, 일정한 직장으로 출근)해야만 결측보정이 가능하다는 점이다.

셋째, 이력 자료가 존재하지 않는 신규 발급된 교통카드의 경우, 보정이 불가능하다.

이러한 몇가지 문제점으로 인해 실제 적용에 있어서는 한계를 가질 수 있을 것으로 보인다. 다음 절에서 제시되는 노선별 승하차 패턴을 이용하는 방법은 이러한 한계를 극복할 수 있다는 점에서 주목할 필요가 있다.

3. 노선별 통행패턴을 고려한 보정

개인별 통행자료를 이용한 방법은 특정한 형태의 개인통행자료를 바탕으로 하기 때문에 결측보정에 한계가 있다. 이러한 개인별 통행이력 자료가 필요없는 노선별 승하차 패턴을 이용하여 보정하는 방법을 제안한다.



〈그림 6〉 노선별 통행패턴을 활용한 보정방안

이러한 방법을 개념화하면 〈그림 6〉에서 나타난 바와 같이 각 노선의 정류장별 승객의 하차 패턴을 분석하여 하차확률에 따라 승차자의 하차 정류장을 결정하는 방법이다. 즉, 완전한 승하차 정보가 있는 데이터를 이용하여 승차 정류장별로 하차정류장을 선택하는 이산확률분포를 작성한다. 그 후 하차정보가 결측된 승객들에 대해 각 하차 확률에 따라 결측보정을 수행하는 방법이다. 이 때 완전한 승하차 정보란, 승·하차시와 환승시 모두 교통카드를 접촉하여 통행의 공간적 패턴이 기록된 통행정보를 의미한다.

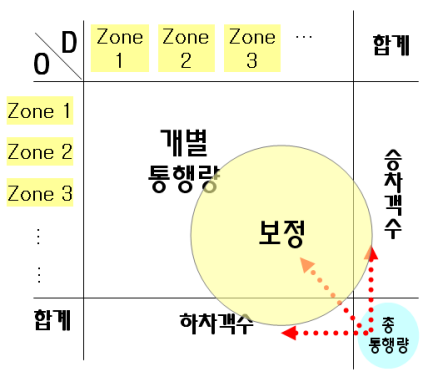
한편, 분석하고자 하는 시간대, 요일, 평일, 주말 등 시계열에 따른 승객의 하차 분포를 각각 따로 작성하여 보정하는 경우 보정의 효과를 증대시킬 수 있다.

4. 총량적 통행수요패턴에 따른 보정

지금까지의 두 방법은 개인의 통행이력이나 노선별 승하차 패턴을 이용한 방법들이다. 즉, 최소한의 통행정보를 기반한 결측보정방안이다. 그러나 이러한 최소한의 정보가 없는 통행자의 경우는 지금까지의 결측보정방법을 활용할 수 없다. 이런 경우의 대표적인 예가 현금승차자의 경우이다. 2006년 말 기준, 하루평균 약 8.6%의 승객이 대중교통 카드가 아닌 현금을 이용한다. 현금승차자 비율은 점차 줄어들고 있으나 아직은 그 비중으로 볼 때 무시할 수 없는 규모이다.

교통카드 데이터를 활용하여 대중교통 OD를 구축하는 경우, 교통카드를 이용한 승객만의 대중교통 OD가 구축되게 된다. 그러나 현금승차자는 노선별로 집계되지 않고 총량으로만 집계되므로 노선별 정보도 활용할 수 없다. 즉, 현금승차자는 OD 및 통행패턴에 대한 아무런 정보가 없으므로 카드데이터의 통행패턴에 기반한 OD에 현금승차자만큼의 통행량에 대해서 총량적 보정을 수행한다.

즉, 현금이용비율을 활용하여 통행량을 산정하고, 산정된 통행량을 이용하여 통행수요 패턴에 따른 O/D쌍별 비율에 따라 분배해 주는 방법을 활용한다.



〈그림 7〉 총량 데이터를 활용한 보정방안

이 방법은 간단하긴 하지만 노선별, 개인별 정보가 전

혀 없기 때문에 결과의 신뢰성을 장담하기 어렵다. 따라서 아무런 정보도 없는 현금승차자 비율에 대해서만 적용하는 것이 합리적일 것이다.

Ⅴ. 결측 보정방법의 적용 및 검증

제안된 보정방안 중 개인별 자료의 이용은 확보된 대중교통 카드데이터 중에서 적절한 이력데이터가 존재하는가가 가장 중요한 문제이므로 별도로 적용성을 검증하는 것은 무의미하다. 또한 세번째 방법인 총량적 통행수요패턴에 따른 보정은 적용방법이 단순하고 그 효과와 한계가 명료하여 별도의 검증이 불필요할 것으로 보인다.

그래서 본 연구에서는 둘째 방법인 "노선별 통행패턴을 고려한 보정방안"에 대해서만 검증하고자 한다.

1. 보정방법의 적용 및 검증 절차

결측 보정방안의 적용 및 검증 절차는 〈그림 8〉과 같이 노선선정, 결측발생, 결측보정, 평가의 과정을 거쳐 수행한다.

〈그림 8〉 노선별 보정방안의 적용 및 평가

1) Step 1. : 분석대상 노선 선정

결측이 적은 4429번(지선버스)의 이용승객들을 중심으로 본 연구의 방법론을 적용 및 검증한다. 우선 4429번 버스의 이용데이터 정보는 다음과 같이 정리할 수 있다.

- 데이터 : 2007년 4월 2일(월) 7시~9시(오전첨두)
- 정류장코드 1번 ~ 10번 (10개 정류소)
- 오전첨두 이용객 : 약 600명
- 원자료의 결측률 : 약 2.7% ~ 4.1%

우선, 검증이 가능한 완전한 데이터를 확보하기 위해 결측 데이터를 제거하여 결측이 없는 완전한 데이터의 형태로 가공하였다.

2) Step 2. : 결측발생

완전한 데이터에서 임의로 결측을 발생시키고, 보정한 후 완전한 데이터와 비교하기 위해 우선 결측 자료가 전혀 존재하지 않는 완전한 데이터에서 임의로 10%, 20%, 30%의 데이터에 대해 결측을 발생시키는 과정을 수행하였다.

3) Step 3. : 결측보정

결측데이터를 보정 및 평가하기 위해 두 가지 방법을 적용하여 비교하였다. 우선 첫번째 방법(1안)은 일률적 보정 방법으로, 결측된 데이터를 제외하여 남겨진 데이터 수를 완전한 데이터 수로 보정하기 위한 계수를 찾아 일괄적으로 각 셀에 곱해 주는 방법이다.

두번째 방법(2안)은 본 연구에서 제안하는 방법으로, 결측된 데이터를 승차정류장별 승객들의 하차정류장 분포확률을 이용하여 각각의 정류장에 배정하여 주는 방법이다.

4) Step 4. : 평가

앞 절의 두 방법을 통하여 보정된 OD 테이블과 완전한 데이터로 구성된 OD 테이블에 대하여 비교·평가한다. 이 때 평가지표로 각 경우의 RMSE와 MAPE 값을 비교하여, 보정 정도에 대해 판단하였다.

$$\mathrm{RMSE} = \sqrt{\frac{(x_i - f_i)^2}{n-1}} \quad (1)$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{x_i - f_i}{x_i}\right| \times 100 \quad (2)$$

여기서, x_i : 실세값 최초 정류장별 OD 실제값
f_i : 추정값, 보정한 추정값
n : 비교 항의 개수

2. 적용 및 평가

본 절에서는 결측 보정방법에 대한 구체적인 적용 및

평가를 수행한다.

우선 첫 번째 단계로 분석대상 노선으로 선정된 4429번 버스의 데이터 중 4월 2일(월)~4월 6일(금)의 데이터를 추출하여 결측 정도에 대해 살펴본다.

〈표 8〉 4429번 지선버스 결측수 (단위 : 통행, %)

구분	4/2 (월)	4/3 (화)	4/4 (수)	4/5 (목)	4/6 (금)
첨두통행량	675	684	633	593	583
첨두결측수	23	23	26	14	16
첨두결측률	3.4%	3.4%	4.1%	2.4%	2.7%
결측제거후	652	661	607	579	567

결측이 없는 완전한 데이터 구축을 위해 원 데이터에서 결측 데이터를 삭제하는 과정을 거친다. 결측 통행 제거 후 구축된 정류장 OD는 다음과 같다.

〈표 9〉 버스 첨두시 실제OD (단위:통행)

	1	2	3	4	5	6	7	8	9	10	계
1	18	28	101	81	109	46	7	0	1	0	391
2	0	0	0	1	0	0	0	0	0	0	1
3	0	0	1	0	0	0	0	0	1	0	2
4	0	0	0	0	0	0	0	0	0	1	1
5	0	0	0	0	0	0	1	0	0	3	4
6	0	0	0	0	0	2	3	5	2	63	75
7	0	0	0	0	0	0	0	1	7	56	64
8	0	0	0	0	0	0	0	1	6	21	28
9	0	0	0	1	0	0	0	0	24	61	86
10	0	0	0	0	0	0	0	0	0	0	0
계	18	28	102	83	109	48	11	7	41	205	652

652통행 중 약 10%인 65통행을 임의로 결측시키고, 결측된 데이터를 일률적 보정방법과 본 연구의 방법에 따라 보정하여 비교한다. 10%의 결측에 대한 분석 이후 20%, 30%에 대한 분석을 차례로 분석한다. 10%의 통행을 임의로 결측발생시킨 결과는 〈표 10〉에 제시된 바와 같다.

〈표 10〉 10% 결측발생 후 정류장 OD (단위:통행)

	1	2	3	4	5	6	7	8	9	10	계
1	16	26	86	73	97	45	6	0	1	0	350
2	0	0	0	1	0	0	0	0	0	0	1
3	0	0	1	0	0	0	0	0	0	0	1
4	0	0	0	0	0	0	0	0	0	1	1
5	0	0	0	0	0	0	1	0	0	3	4
6	0	0	0	0	0	2	3	4	2	57	68
7	0	0	0	0	0	0	0	1	7	48	56
8	0	0	0	0	0	0	0	1	6	21	28
9	0	0	0	1	0	0	0	0	21	56	78
10	0	0	0	0	0	0	0	0	0	0	0
계	16	26	87	75	97	47	10	6	37	186	587

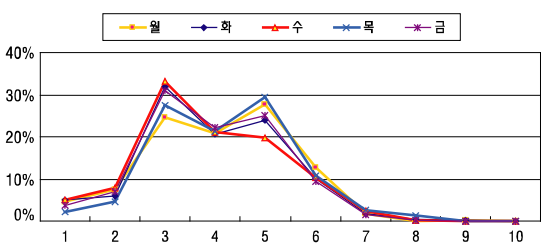
우선 일률적 보정방안인 "1안"에 따라서 보정 OD를 구축한다. 652 통행 중 65통행을 결측시켰기 때문에 데이터 수는 587 통행인데, 587 통행을 652 통행으로 보정하기 위한 계수 "1.11"을 모든 셀에 곱하여 결과를 산출한다. 〈표 11〉은 "1안"을 통해서 보정된 테이블이다.

〈표 11〉 "1안"보정 후 OD (단위:통행)

	1	2	3	4	5	6	7	8	9	10	계
1	18	29	96	81	108	50	7	0	1	0	390
2		0	0	1	0	0	0	0	0	0	1
3			1	0	0	0	0	0	0	0	1
4				0	0	0	0	0	0	1	1
5					0	0	1	0	0	3	4
6						2	3	4	2	63	74
7							0	1	8	53	62
8								1	7	23	31
9									23	62	85
10										0	0
계	18	29	97	82	108	52	11	6	41	205	649

위의 과정을 통해 계산된 RMSE 값은 1.06, MAPE는 6.44로 계산되었다.

다음은 "2안"을 통해서 보정하는 과정이다. 우선, 승차정류장별 승객들의 하차정류장 분포를 분석하였다. 1번 정류장에서 승차한 승객들을 대상으로 하차정류장 분포도를 작성한 그림은 〈그림 9〉와 같다.



〈그림 9〉 1번 정류장 승객의 하차 분포

4월 2일(월)~6일(금) 동안의 평균분포도를 산출하였다. 이 하차정류장 분포는 하차정류장에 대한 선택확률분포가 되는 것이다. 4월 2일(월)의 분포는 임의 결측통행을 제거한 데이터를 이용하여 분포를 구하였다.

1번 정류장에서 승차했으나 결측이 발생한 통행은 "41

통행"으로 〈그림 9〉에 나타난 하차정류장의 확률분포에 따라 하차 정류장을 배정한다. 따라서, 개별 정류장의 하차 승객수는 결측 발생횟수 41회에 각각 정류장별 하차 확률을 곱하여 산정한다. 즉, 각 정류잘별 하차 인원은 X*P(X)이다. 이렇게 각 승차정류장별 하차비율을 하나의 이산확률분포로 작성하였다. 그 결과 〈표 12〉와 같다.

〈표 12〉 "2안"10% 결측발생 후 OD (단위:통행)

	1	2	3	4	5	6	7	8	9	10	계
1	4.3	6.6	29.8	21.2	25.0	10.7	2.0	0.5	0.1	0.0	100
2		0.0	0.0	50.0	0.0	0.0	50.0	0.0	0.0	0.0	100
3			21.4	21.4	7.1	7.1	14.3	0.0	28.6	0.0	100
4				0.0	0.0	20.0	20.0	0.0	0.0	60.0	100
5					0.0	0.0	9.1	0.0	0.0	90.9	100
6						0.9	4.2	6.0	9.3	79.7	100
7							2.1	1.4	8.6	88.0	100
8								1.9	20.9	77.2	100
9									22.1	77.7	100
계	2.50	3.87	17.59	12.66	14.73	6.53	2.17	1.17	6.13	32.66	100

이와 동일한 방법으로 . 이러한 과정을 "3번 정류장", "4번 정류장", "5번" …… 에 적용하여 결측이 발생한 모든 통행을 보정하여 배정하였다. 〈표 13〉은 "2안"을 통해서 보정이 완료된 테이블이다.

〈표 13〉 "2안" 보정OD (단위 : 통행)

	1	2	3	4	5	6	7	8	9	10	계
1	18	29	98	82	107	49	7	0	1	0	391
2	0	0	0	1	0	0	0	0	0	0	1
3	0	0	1	0	0	0	0	0	0	0	1
4	0	0	0	0	0	0	0	0	0	1	1
5	0	0	0	0	0	0	1	0	0	3	4
6	0	0	0	0	0	2	3	4	3	63	75
7	0	0	0	0	0	0	0	1	8	55	64
8	0	0	0	0	0	0	0	1	6	21	28
9	0	0	0	1	0	0	0	0	23	62	86
계	18	29	99	84	107	51	11	6	41	205	651

"2안"의 RMSE 값이 0.75로 "1안"의 RMSE 값 1.06보다 작고, MAPE 또한 각각 1.96와 6.44로 2안이 1안보다 나은 결과를 보여주고 있다.

지금까지의 과정을 결측률이 20%, 30%로 확장된 경우로 가정하여 추가로 분석하였다. 분석 결과, 각 대안별 RMSE와 MAPE값은 〈표 14〉와 같다.

〈표 14〉 평가 결과 요약

구분		10% 결측	20% 결측	30% 결측
1안	RMSE	1.06	1.70	1.67
	MAPE	6.44	16.41	33.84
2안	RMSE	0.75	1.32	1.14
	MAPE	1.96	5.23	10.06

평가 결과에서 나타난 바와 같이 결측률에 상관없이 본 연구에서 제시한 결측보정방법이 효과가 있는 것을 알 수 있다.

3. 결측보정의 활용 방안

본 연구에서 제시한 결측보정은 향후 대중교통 카드 데이터 기반의 대중교통 OD분석에 있어서 기본적인 절차가 될 것이다. 그러므로 본 연구에서 제시한 3가지 방법은 각 방법의 특성을 파악하여 활용할 필요가 있다.

우선, 첫 번째 개인이력정보 활용방법은 각 개인의 통행을 분석하여 보정하므로 정확한 보정은 가능하나 개인 통행분석 및 패턴 발견이 용이하지 않으므로 적용상에 한계가 있다.

두번째, 노선별 승하차 패턴을 이용하는 방법은 첫 번째 방법에 비해서는 떨어질 수 있으나 일정수준이상의 보정 정확성을 유지하면서 노선별 특성을 유지할 수 있다는 점에서 유용하다고 할 수 있다. 단, 각 노선의 정류장별 하차 확률분포를 산출하기 위한 과정이 요구되긴 하지만 첫 번째 방법에 비해서는 용이한 적용이 가능하다.

세 번째 방법은 적용은 단순하고 용이하나 그 결과의 신뢰성에 대해서는 검증이 어려우므로 현금승차자와 같은 기타의 정보가 전혀 없는 통행에 대해서만 필요할 경우 적용하여야 할 것이다.

이러한 특성을 고려할 때 보다 정확한 O/D가 필요한 경우는 세 방법을 단계적으로 활용해야 할 것이다. 그러나 대략적인 패턴을 분석할 경우, 두 번째 방법만으로 전체 데이터를 보정하는 것도 유용한 방법이 될 것이다. 분석 목적에 따라 각 방법을 적절히 활용하는 지혜가 필요하다.

Ⅵ. 결론

대중교통 교통카드 도입 이후, 점차 이용이 증가되고 있다. 이러한 현상은 이용자의 편의를 제공해주는 면에

서도 중요하지만 카드 데이터를 통해 얻을 수 있는 자료를 고려할 때 교통분석 및 정책적 측면에서 중요한 의미를 가지고 있다. 그 중에서 특히 존별 대중교통 통행수요(O/D)를 손쉽게 파악할 수 있다는 점에서 의의가 크다.

본 연구에서는 카드데이터를 통해 대중교통 존별 통행수요(O/D)를 파악함에 있어서 반드시 선행되어야 하나 아직 연구사례가 없었던 카드데이터의 오류와 결측에 관해 살펴보았다. 그 과정에서 통행수요(O/D)분석과 관련한 오류나 결측에 대한 인식을 제공하였고, 결측에 대한 보정방안을 제안하였다. 그리고 제시된 결측방안들에 대한 적용 및 평가와 함께 활용방안을 제시하였다.

본 연구에서는 향후 신뢰도나 활용도 측면에서 대단히 중요한 역할을 하게 될 대중교통 카드데이터의 이용을 위한 기본적인 데이터 검증 및 보완 절차를 마련했다는 점에서 연구의 의의를 찾을 수 있다.

그러나 본 연구에서 살펴본 여러 가지 카드데이터의 오류는 앞으로도 보다 많은 데이터의 검토가 수반되어야 하고, 결측의 보정방안들은 현장 적용과정에서 많은 개선이 필요할 것이다. 특히, 보다 많은 노선들에 본 연구에서 제안한 방법론의 적용을 통한 검증이 필요할 것으로 판단되며, 이는 본 연구의 한계이자 향후 연구과제로 돌리고자 한다.

또한, 본 연구에서는 개인별·노선별 통행패턴 방법을 따로 적용하되, 개인통행 데이터의 특성과 한계를 인식하여 각 방법론을 별도의 단계로 구분하여 순차적으로 적용하는 방안을 제시하였다. 향후 본 연구에서 제시한 방법론들을 결합하여 보다 높은 신뢰성과 적용성을 기대할 수 있는 방안모색을 향후 연구과제로 남긴다.

향후 대중교통 정책의 가장 중요한 기반이 될 카드데이터의 신뢰성과 활용성이 확보될 때 도시의 교통문제 해결의 실마리가 마련될 것으로 기대한다.

참고문헌

1. 건설교통부(2006), "대중교통기본계획 수립".
2. 김순관(2005), "서울시 OD조사 신뢰성 증대방안 연구", 서울시정개발연구원.
3. 김현석(2006), "순환확률분포를 이용한 교통량 결측자료 보정 모형에 관한 연구", 서울대학교 박사학위 논문.
4. 박진영(2006), "대중교통 정책수립을 위한 교통카드자료 활용방안", 한국교통연구원.
5. Donald B. Rubin(2004), "Multiple imputation for non-response in surveys", A john wiley & sons.

♧ 주 작 성 자 : 박준환
♧ 교 신 저 자 : 박준환
♧ 논문투고일 : 2007. 8. 20
♧ 논문심사일 : 2007. 11. 29 (1차)
2008. 3. 18 (2차)
♧ 심사판정일 : 2008. 3. 18
♧ 반론접수기한 : 2008. 8. 31
♧ 3인 익명 심사필
♧ 1인 abstract 교정필